WPS 3294

Imputed Welfare Estimates in Regression Analysis[1]

Chris Elbers[2] Jean O. Lanjouw[3] Peter Lanjouw[4]

April 26, 2004

[1]We thank Ravi Kanbur, Tony Venables and other participants at the WIDER project meeting on Spatial Inequality in Development, May 2003, for comments on an earlier draft of this paper. Also we wish to thank Jishnu Das, Elisabeth Sadoulet and Alain de Janvry for valuable input, and François Bourguignon and Martin Ravallion for stimulating our interest in the questions pursued here. Finally, the paper has benefited from the comments of two anonymous referees.

[2]Amsterdam Institute for International Development; and Vrije Universiteit Amsterdam, celbers@feweb.vu.nl.

[3]ARE Department, U.C. Berkeley, Brookings Institution and the Center for Global Development, Washington, DC, jlanjouw@are.berkeley.edu.

[4]World Bank, Washington, DC, planjouw@worldbank.org. Financial support was gratefully received from the Bank Netherlands Partnership Program. None of the views expressed here should be taken to represent those of the World Bank or affiliated organizations.

Abstract

We discuss the use of imputed data in regression analysis, in particular the use of highly disaggregated welfare indicators (from so-called "poverty maps"). We show that such indicators can be used both as explanatory variables on the right-hand side and as the phenomenon to explain on the left-hand side. We try out practical ways of adjusting standard errors of the regression coefficients to reflect the error introduced by using imputed, rather than actual, welfare indicators. These are illustrated by regression experiments based on data from Ecuador. For regressions with imputed variables on the left-hand side, we argue that essentially the same aggregate relationships would be found with either actual or imputed variables. We address the methodological question of how to interpret aggregate relationships found in such regressions.

Introduction

The growing access of researchers to household data makes possible the estimation of inequality and poverty measures at very disaggregated levels. In Elbers, Lanjouw and Lanjouw (2003) we describe a procedure that combines the broad coverage of a census or large survey and the detail of household survey data to arrive at estimators that are quite precise - comparing very favorably to estimates based on either source alone. Using this strategy, estimates of local welfare (so-called poverty maps) have been constructed for many countries (see Demombynes, *et al.*, 2003, for examples). These maps provide useful information about the geographic spread of relative poverty and inequality that can be directly useful to policy makers pursuing poverty alleviation or development goals.

In addition to their direct informational use, the imputed welfare estimates also provide a wealth of distributional information that could be used in economic analysis. Theories abound regarding what causes localities to be poor or unequal and how these characteristics might affect other social or economic outcomes. In the absence of appropriate data, it has been difficult to explore these ideas empirically. Imputed welfare estimates could enable more extensive applied distributional analysis. In such studies, however, it will be important to take account of the fact that the estimates are exactly that, estimates, and not data. Thus, in this paper we discuss the econometric issues raised when using imputed welfare estimates in regression analysis - as either a dependent variable or an explanatory variable.

Most of the issues we discuss are quite general and arise in all situations using predicted variables. An extensive discussion, for example, can be found in Murphy and Topel (1985). We focus here on the use of imputed welfare variables. We make suggestions regarding some particular problems that might arise when using these variables, and explore the importance of various issues using Ecuador as an illustration.

Specifically, we are interested analyzing the relationships between a true welfare measure, W, and other variables in what we will call "downstream" regressions. W is unknown but we have consistent estimates of the expected value of W denoted by $\widetilde{\mu}$. The estimate $\widetilde{\mu}$ is an error-ridden variable. However, by its construction it can be understood as an instrumented version of W and standard results, including consistency, for IV estimators obtain. Although the welfare estimates are more complex than standard instrumental variables, we show how one can use information about the distribution of $\widetilde{\mu}$ to calculate consistent standard errors for the downstream coefficient estimates.

Using an imputed value to serve as an explanatory variable may create an endogeneity problem if the variables used in its construction are correlated with the disturbance in the downstream regression. We examine the likely importance of this concern when the correlated variables and the regressions are at various levels of aggregation and suggest ways to avoid

introducing an endogeneity bias. The use of imputed values may also *resolve* an endogeneity problem. In some situations the true value of W may be correlated with the disturbance term. In this case, one would like to instrument W using variables uncorrelated with the disturbance in the downstream regression. With attention paid to how they are constructed, predicted values $\widetilde{\mu}$ can be interpreted as useful instruments for W when W is endogeneous. That is, predicted values can be superior to the unknown true values W.

The construction of the imputed welfare estimates is briefly described in Section 1. In Section 2 we discuss the use of these estimates as an explanatory variable and in Section 3 describe practical approaches to calculating consistent standard errors for the downstream regression coefficients. Endogeneity issues are considered in Section 4. In Section 5 we discuss the use of an imputed value as the dependent variable in the downstream regression. The last section concludes.

1 Calculation of Imputed Welfare Estimates

Denote by W a measure of poverty or inequality based on the distribution of a household-level variable of interest, y_h, for instance per-capita expenditure. Data on y_h as well as a number of covariates $\mathbf{z}_h$ are available from a household survey, where h refers to a household included in the survey and bold variables indicate vectors and matrices. Measuring household per-capita expenditure reliably is very costly, therefore this kind of survey is typically only representative at high levels of aggregation, say the province level. Consequently, welfare indicators W, based on direct observations of y, are also at best available at the province level. By bringing in information from other data sources we can overcome this limit and compile welfare estimates at levels of aggregation far below the province level.

The idea is that from the household survey we can estimate the joint distribution of y_h and one or more of the covariates z_{hi}. Assume that a larger-scale sample or a census of households is available besides the survey and containing observations of some of the components of $\mathbf{z}_h$.[1] By estimating the joint distribution of y_h and the subset of covariates also in the census,[2] this estimated distribution can be used to generate the distribution of y_h for any sub-population in the larger sample conditional on the sub-population's observed characteristics. This, in turn, allows us to generate the conditional distribution of W for sub-populations. We do this by means of simulation.

In what follows we let $\mathbf{z}$ denote the vector of covariates which can be linked to both survey

[1] Ideally survey and census would refer to the same year. If not, it is necessary either to assume that the relationship between consumption and observables did not change over the period between the data souces (as we do below), or the model estimated must be extended to capture any change.

[2] Actually, we can do better than that. We can bring in any variable that can be linked both to survey and census households. In practice this appears to be a crucial improvement. See Elbers *et al.* (2003) for details.

and census households. The first step, which we call the "first stage", is to develop an accurate empirical model of y_{ch}, the per-capita expenditure of household h in sample cluster c. Typical applications have used a log-linear approximation to the conditional distribution of y_{ch},

$$\ln y_{ch} = E[\ln y_{ch}|\mathbf{z}_{ch}] + u_{ch} \approx \mathbf{z}'_{ch}\boldsymbol{\gamma} + \eta_c + \varepsilon_{ch}. \tag{1}$$

By including cluster random effects η_c in the equation we allow for a within cluster correlation of disturbances u_{ch}. The error components η and ε are assumed to be independent of each other. They are uncorrelated with observables, $\mathbf{z}_{ch}$, by construction.

Suppose that there are M households in a target population and household h has m_h family members. In general one will want to account for household size in welfare measures, so we write $W(\mathbf{m}, \mathbf{Z}, \boldsymbol{\gamma}, \mathbf{u})$, where $\mathbf{m}$, $\mathbf{Z}$ and $\mathbf{u}$ are conformable arrays of household size, observable characteristics and disturbances, respectively. The expected value of W given the observable characteristics and the model of expenditure is denoted $\mu = E[W|\mathbf{m}, \mathbf{Z}, \boldsymbol{\zeta}]$, where $\boldsymbol{\zeta}$ is the vector of model parameters, including $\boldsymbol{\gamma}$ and any parameters describing the distribution of the disturbances η and ε.

In the second stage construction of our estimator of μ, we replace $\boldsymbol{\zeta}$ with consistent estimators, $\widehat{\boldsymbol{\zeta}}$, from the first stage expenditure regression. Simulation is used to obtain $\widetilde{\mu}$ =E{E[W | $m, Z,\widehat{\zeta}$]}, where the outer expectation is over the sampling distribution of $\hat{\zeta}$, given $\zeta = \hat{\zeta}$.

The difference between $\widetilde{\mu}$, the estimator of the expected value of W for a population, and the actual level may be written

$$\xi = W - \widetilde{\mu} = (W - \mu) + (\mu - \widetilde{\mu}). \tag{2}$$

Thus the prediction error ξ has two components:[3]

Idiosyncratic Error - $(W - \mu)$. The actual value of the welfare indicator for a population deviates from its expected value, μ, as a result of the realizations of the unobserved component of expenditure. This component increases as one focuses on smaller target populations.

Model Error - $(\mu - \widetilde{\mu})$. This component of the prediction error is determined by the properties of the first-stage estimators so it does not increase or fall systematically as the size of the target population changes.

Elbers, Lanjouw and Lanjouw (2003) show that these error components are asymptotically normal, converging in the population size M and the household survey size s. In typical applications the overlap between the target population and the survey is virtually nil, so the

[3]There is also simulation error, but we will assume that it has been made small enough to ignore.

variance of the total prediction error is the sum of individual error variance components:

$$\mathrm{V} = \mathrm{V_I} + \mathrm{V_M}. \tag{3}$$

2 Predicted Welfare as an Explanatory Variable

Consider first using imputed welfare measures on the "right-hand side" of a regression. Start from the general regression equation

$$D = \mathbf{x}'\boldsymbol{\alpha} + W\beta + \tau. \tag{4}$$

D is explained by regressor vector $\mathbf{x}$ and welfare indicator W. We are interested in estimating β, the effect of W on D. In our case W is not observed, so we use estimates of expected welfare $\widetilde{\mu}$. As discussed above, our predicted welfare is related to W as

$$W = \widetilde{\mu} + \xi. \tag{5}$$

Substituting this equation in the regression equation one gets

$$D = \mathbf{x}'\boldsymbol{\alpha} + \widetilde{\mu}\beta + (\xi\beta + \tau). \tag{6}$$

It follows that β can be consistently estimated if $\mathbf{x}$ and $\widetilde{\mu}$ are uncorrelated with ξ and τ, or if $\widetilde{\mu}$ is uncorrelated with $\mathbf{x}$, ξ, and τ.

In our case, $\widetilde{\mu}$ is a consistent estimator of the conditional expectation of W, ξ is a prediction error and so $\widetilde{\mu}$ and ξ are uncorrelated. Thus, if the other standard properties are met, using $\widetilde{\mu}$ in a regression rather than W still yields consistent estimates.

Note the importance of the fact that $\widetilde{\mu}$ is a prediction. If instead $\widetilde{\mu}$ were some other proxy for W, then the error ξ would represent a form of measurement error which is correlated with $\widetilde{\mu}$. It is possible that an analyst might inadvertently introduce measurement error into the regression when W is a discrete variable (say, poverty status at the household level). Because its expectation, $\widetilde{\mu} = E(W|\mathbf{z})$, is a continuous variable (expected poverty status given household characteristics $\mathbf{z}$) it is tempting in this circumstance to use a discrete version of $\widetilde{\mu}$, say $\hat{W}$, in equation (6) (i.e. setting $\hat{W} = 1$ for a household if $\widetilde{\mu} \geq 0.5$ and 0 otherwise). This would not be advisable because measurement error typically leads to attenuation bias in the estimation of β.

2.1 Standard errors on downstream regression coefficients

When using imputed welfare as an explanatory variable in a regression equation, the estimated standard errors on the regression coefficients must take account of additional noise in the estimates. To see this, insert equation (2) into regression equation (4) to obtain

$$D = \mathbf{x}'\boldsymbol{\alpha} + \widetilde{\mu}\beta + (\mu - \widetilde{\mu})\beta + (W - \mu)\beta + \tau. \tag{7}$$

The error term ψ in this regression consists of the following components

$$\psi = (\mu - \widetilde{\mu})\beta + (W - \mu)\beta + \tau. \tag{8}$$

Thus there are three sources of error in the estimates of the downstream regression coefficients of equation (7). One is the standard sampling error (represented by τ), a second derives from the difference between W and the true expectation μ (the idiosyncratic error), and the third is from the difference between μ and its estimate $\widetilde{\mu}$ (the model error). Except for the idiosyncratic error part $(W - \mu)\beta$ this error decomposition is very similar to formula (8) in Murphy and Topel (1985). As in their case, estimating equation (7) directly from data on D, $\mathbf{x}$, and $\widetilde{\mu}$, would typically underestimate the true variance of the estimator for β, because the model error term $(\mu - \widetilde{\mu})\beta$ creates correlation across the observations.

The source of correlation is clear. Recall that model error arises because computation of μ requires knowledge of $\boldsymbol{\zeta}$, the parameter vector that describes consumption. As explained in section 1, estimates of these parameters are used to impute (the conditional distribution of) consumption expenditure for all households in a target population. Because the same expenditure model is applied to a group of households the same (erroneous) parameter estimates are applied to all of them, thus creating correlation across errors in the prediction of those households' consumption expenditure. This is *un*like correlation resulting from location effects in that correlation due to model error is very likely to carry over to higher levels of aggregation (sub-district, district, and so on).[4]

It is nonetheless straightforward to estimate the full variance of the estimated parameters $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\beta}$ of the regression equation (7) once the correlation of the model error across downstream observations is known. To see this, rewrite the regression equation as

$$\mathbf{D} = \mathbf{X}\boldsymbol{\lambda} + (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}})\beta + \mathbf{e},$$

where $\mathbf{X}$ is the matrix of observations $(\mathbf{x}, \widetilde{\mu})$, $\boldsymbol{\lambda} = (\boldsymbol{\alpha}, \beta)$ is the vector of regression parameters, and $\mathbf{e} = (\mathbf{W} - \boldsymbol{\mu})\beta + \boldsymbol{\tau}$ is the residual part not related to model error. $\boldsymbol{\Sigma}_{\mathrm{M}}$ is the covariance

[4] Typically it is possible to estimate separate consumption models at the level of strata. Estimates for sub-populations belonging to different survey strata then do not have correlated model error.

matrix of model error in $\widetilde{\boldsymbol{\mu}}$. If the components of $\mathbf{e}$ are i.i.d. and there are no endogeneity issues plaguing the regression, then OLS is consistent and the OLS estimator for $\boldsymbol{\lambda}$ has (asymptotic) variance

$$\mathrm{Var}(\widehat{\boldsymbol{\lambda}}) = \sigma_e^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} + \beta^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\boldsymbol{\Sigma}_{\mathrm{M}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}. \tag{9}$$

Alternatively, feasible GLS could be used instead of OLS. For clarity of exposition we do not discuss this here. GLS would often be the preferred method if downstream regression equation (4) is estimated at the household level. This is because households within the same cluster typically share a common location effect which affects their consumption level in a similar way (the disturbance component η_c in equation 1). Thus the idiosyncratic part of the prediction error (W - μ)β is correlated when observations are at the level of households. This complication does not occur if the regression is at higher levels of aggregation.[5]

Below we will refer to the first term in (9) as the "sampling" part of the variance and the second part as the "model" part of the variance. In the next section we try out alternative ways to compute $\mathrm{Var}(\widehat{\boldsymbol{\lambda}})$.

3 Estimation of Standard Errors

Suppose, first, that one knew the true expected value of W, that is $\widetilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{\mathrm{M}} = 0$. In this case, the second term of $\mathrm{Var}(\widehat{\boldsymbol{\lambda}})$ in equation (9) would disappear. The downstream regression model (7) and standard errors could be estimated in the usual way.

If OLS is used in a household-level regression, the sampling part of the variance in $\widehat{\boldsymbol{\lambda}}$ can still be estimated consistently using standard methods. One approach is to estimate the model with $\widetilde{\mu}$ and then use downstream regression residuals and a robust variance formula (see, for example, Greene (2000), equation 11-14). With large numbers of downstream observations, however, this is cumbersome. Alternatively one could bootstrap the variance by resampling out of the downstream data (including $\widetilde{\mu}$), re-estimating the model many times, and calculating the variance of the resulting estimates of $\boldsymbol{\lambda}$.[6] By bootstrapping, any correlation in the idiosyncratic error across observations due to location effects is incorporated in the variance estimation directly. Note again that, in general, if the downstream regression is estimated at a level of aggregation higher than the level of any location effects then the prediction errors (W-μ) are no longer correlated and these steps are unnecessary.

We now turn to estimation of the second term in (9), the variance due to model error in the imputed welfare estimates. From our earlier discussion, recall that $\boldsymbol{\zeta}$ denotes the true parameter

[5]The prediction errors might also be heteroscedastic, although in our experience the variance of $\widetilde{\mu}$ does not appear systematically related to its size, as one might perhaps expect.

[6]The bootstrapping should be nested like the error structure in equation (1) by first drawing groups of households at the level of aggregation where η_c applies, and then drawing households randomly within the group.

vector underlying expenditure equation (1), and write $\mu = \mu(\boldsymbol{\zeta})$ to stress the dependency of μ on $\boldsymbol{\zeta}$. Following Murphy and Topel we could relate the model error $\mu - \widetilde{\mu}$ to the error in $\widehat{\boldsymbol{\zeta}}$ by linear approximation:

$$\mu - \widetilde{\mu} \approx \frac{\partial \mu}{\partial \boldsymbol{\zeta}}(\widehat{\boldsymbol{\zeta}})(\boldsymbol{\zeta} - \widehat{\boldsymbol{\zeta}})$$

and use the estimated variance of $\widehat{\boldsymbol{\zeta}}$ to infer the (asymptotic) error distribution of $\widetilde{\mu}$.[7] However, for the purpose of calculating the variance in downstream regression coefficients the simplest approach is to simulate the distribution of $\mu - \widetilde{\mu}$ directly (see below, section 3.1). Bypassing the calculation of derivatives has the additional advantage that (small-sample) bias arising from the linear approximation in Murphy and Topel's approach is avoided.

The simulations are described in the following subsection. They are done under the assumption that there is no correlation between the model error part $(\mu - \widetilde{\mu})\beta$ and the other error components of equation (8), $(W - \mu)\beta + \tau$.[8] The main justification for this assumption is that the model error $\mu - \widetilde{\mu}$ is ultimately caused by sampling variation in the survey from which $\boldsymbol{\zeta}$ is estimated. This survey typically covers only a tiny fraction of the population for which the welfare indicators $\widetilde{\mu}$ have been compiled[9] and may come from a different time period if census and survey—regrettably—are from different years.

To perform the calculations described in subsection 3.1 below one needs to employ the data that were used in the computation of the estimators $\widetilde{\mu}$. Since most researchers will not have access to the unit record level data, particularly not for a census, we propose in subsection 3.2 several alternative ways to approximate $\boldsymbol{\Sigma}_{\mathrm{M}}$ when pieces of information are unknown. Ultimately our goal is to find a parsimonious and satisfactory representation of $\boldsymbol{\Sigma}_{\mathrm{M}}$ which could be reported together with the welfare estimates in a poverty mapping project so that analysts can readily adjust standard errors from regressions involving imputed welfare estimates. In the final subsection we give empirical illustrations for Ecuador.

3.1 Estimation with unit record data

The model error part in the variance of $\widehat{\boldsymbol{\lambda}}$, i.e. the second term in (9), is due to error in the consumption model used to estimate μ. The combined effect of sampling and model error can be simulated by drawing from the distribution of $\widetilde{\boldsymbol{\mu}}$ and $\mathbf{e}$ and re-estimating the downstream regression model. As discussed in Section 1, the estimates $\widetilde{\boldsymbol{\mu}}$ are determined by household survey data and a vector of estimated consumption model parameters $\widehat{\boldsymbol{\zeta}}$. In the simulations we take the following steps:

[7] See Murphy and Topel (1985), page 374. This approach is taken in Elbers, Lanjouw and Lanjouw (2003).

[8] In the terminology of Murphy and Topel this is the case of *independent random components*.

[9] Dependence would be completely eliminated if surveyed households could be excluded from the census data. However, identifying survey households in the census is practically impossible.

1. Draw vectors $\widehat{\boldsymbol{\zeta}}^r$, $r = 1, ..., R$, from the appropriate sampling distribution (see Elbers, *et al.*, 2003, for this distribution).

2. Draw a simulated vector of downstream regression disturbances $\mathbf{e}^r$ from an estimated distribution of $\mathbf{e}$. Construct a new vector of simulated dependent variables

$$\mathbf{D}^r = \mathbf{x}\widehat{\alpha} + \widetilde{\mu}\widehat{\beta} + \mathbf{e}^r.$$

3. Simulate the expected welfare measures implied by each, $\widetilde{\boldsymbol{\mu}}^r = \widetilde{\boldsymbol{\mu}}(\widehat{\boldsymbol{\zeta}}^r)$. (The covariance matrix of the $\widetilde{\boldsymbol{\mu}}^r$ is $\widehat{\boldsymbol{\Sigma}}_{\mathrm{M}}$, which however is not needed in this more direct procedure.)

4. Estimate the downstream regression coeffient $\boldsymbol{\lambda}$ using the simulated $\mathbf{D}^r$ and $\mathbf{X}^r$, where $\mathbf{X}^r$ is the matrix of observations $(\mathbf{x}, \widetilde{\mu}^r)$. Note that $\widetilde{\boldsymbol{\mu}}$, not $\widetilde{\boldsymbol{\mu}}^r$ is used in step 3 above.

The variance of these R simulated values $\widehat{\boldsymbol{\lambda}}^r$ gives an estimate of the total error variance of $\widehat{\boldsymbol{\lambda}}$.[10]

3.2 Estimation without unit record data

The estimation strategy described in the preceding subsection requires access to the unit record data. Note, however, that it is quite straightforward to report the model variance for each $\widetilde{\mu}$. If the $\widetilde{\mu}$ were independent across observations this simulation could be done at the level of the $\widetilde{\mu}$ without need for access to the data used in the construction of $\widetilde{\mu}$. However, as discussed earlier, the estimates of μ will often be correlated. For example, typically one consumption model is estimated for rural areas and another for urban areas. Then welfare estimates imputed for rural populations (households, villages, sub-districts, etc.) in the downstream data would share model error, and likewise for the urban populations. It is not easy, however, to characterize this correlation because it is dependent on the values for the z variables associated with any pair of downstream observations (households, villages, sub-districts, etc.). It becomes yet harder to characterize if the unit of observation in the downstream regression mixes households having different consumption models. Thus, the straightforward approach - when it is possible - is to begin the simulation from the estimated consumption parameters, $\widehat{\boldsymbol{\zeta}}$, as above.

If the unit record data are unavailable, we must start from the $\widetilde{\mu}$ and use some approximation to their correlation. Take the typical case where a different consumption model is estimated for each stratum in the household survey data. There is then no correlation between households in different strata. Let $-1 \leq K_s \leq 1$ be the correlation coefficient between units within stratum s. For example, suppose $\widetilde{\boldsymbol{\mu}}$ is a vector of estimates for four households, two in stratum F and

[10]The estimator $\widehat{\boldsymbol{\lambda}}$ is consistent but biased when μ is unknown. The average of the simulated coefficient estimates $\widehat{\boldsymbol{\lambda}}^r$ derived from this procedure would give an unbiased estimator under the (estimated) sampling distribution of $\widetilde{\boldsymbol{\mu}}$.

two in stratum Q. $\mathrm{V}_{\mathrm{M}h}$ represents the model part of the variance of $\widetilde{\mu}_h$ for household $h = 1,...,4$ (see equation 3). Then model covariance matrix for $\widetilde{\boldsymbol{\mu}}$ is:

$$\boldsymbol{\Sigma}_{\mathrm{M}} = \begin{bmatrix} \mathrm{V}_{\mathrm{M1}} & K_F\sqrt{\mathrm{V}_{\mathrm{M1}}\mathrm{V}_{\mathrm{M2}}} & 0 & 0 \\ & \mathrm{V}_{\mathrm{M2}} & 0 & 0 \\ & & \mathrm{V}_{\mathrm{M3}} & K_Q\sqrt{\mathrm{V}_{\mathrm{M3}}\mathrm{V}_{\mathrm{M4}}} \\ & & & \mathrm{V}_{\mathrm{M4}} \end{bmatrix}. \tag{10}$$

We explore a number of different approximations for the matrix $\boldsymbol{\Sigma}_{\mathrm{M}}$. The purpose is to give guidance to downstream researchers whose information about the true matrix may be limited. It is also to suggest the type of information that should be provided by those producing welfare estimates to improve their usefulness. Each approximation yields an estimated matrix $\widehat{\boldsymbol{\Sigma}}_{\mathrm{M}}$.

It is likely that the downstream researcher has little or no information about the differing degrees of correlation in model error across units. Thus we try to approximate these values with correlation coefficients that are constant within a given stratum (the K_s). If the welfare estimates are coming from secondary sources, the researcher also may know only the total variance in $\widetilde{\mu}$, V, and not the portion due to model error. In this case a second approximation is needed, with $\widehat{\mathrm{V}}_{\mathrm{M}} = G\mathrm{V}$. Reasonable values for K and G will depend on the level of aggregation of the $\widetilde{\mu}$.

In the following subsection we present examples from Ecuador. These show the importance of including the model part of $\mathrm{Var}(\widehat{\boldsymbol{\lambda}})$, and indicate how sensitive estimates of the variance are to assumptions about the degree of correlation in the imputed welfare estimates, $\widetilde{\mu}$. As will become clear from that discussion, the approximations outlined above do not perform particularly well. One might try to obtain a reasonable upperbound for the variance, replacing $\boldsymbol{\Sigma}_{\mathrm{M}}$ in equation (9) by a diagonal matrix $\rho\mathbf{I}$ for sufficiently high ρ. The model error variance part then simply becomes

$$\rho\beta^2(\mathbf{X}'\mathbf{X})^{-1}.$$

The question is, of course, what an appropriate value for ρ would be. We have used the maximum total variance, V, found among welfare estimates at the level used in the downstream regression.

Finally, one way to assess the possibilities for a parsimonious representation of $\boldsymbol{\Sigma}_{\mathrm{M}}$ is to see how many terms in a singular value decomposition of it are needed. Results are summarized in Tables 1 and 2 below.

3.3 Experiments for Ecuador

Our empirical examples use data from Ecuador. Expected welfare is based on household per-capita expenditure. Consumption models are estimated using the 1994 Ecuadorian *Encuesta*

Sobre Las Condiciones de Vida, a household survey following the general format of a World Bank Living Standards Measurement Survey. It is stratified by eight regions and separate models are estimated for each stratum. We were able to capture most of the effect of location on consumption with available explanatory variables. This means that there is little correlation across households in their idiosyncratic error. The models are used to impute welfare measures for target populations in the 1990 Ecuadorian census. (See Elbers, Lanjouw, and Lanjouw, 2002, for a full discussion of the estimation procedure and diagnostics.)

We study canton-level regressions where the dependent variable is "garbage", the percentage of households in the canton whose garbage is collected by the municipal trucks.[11] The explanatory variables are a normalized measure of cantonal population size and a point estimate of welfare, either the local headcount or the local inequality index GE(0.5). Moreover, province dummies have been added to avoid obvious omitted variables bias. The estimations use pooled data for the Rural Costa and Sierra regions for a total of 164 cantons with an average population of 26,650.

The regression results are reported at the top of Tables 1 and 2, respectively, where the reported standard errors reported in the top panel of the table include only the sampling part of the error. Local poverty is associated with a lower incidence of garbage collection, while greater community inequality is associated with a higher level of service. Both regressions have reasonable R^2s, given that these are cross-section regressions. The coefficients on the province dummies are not reported. Each of these dummies is highly significant in both regressions. On the other hand, without the dummies the parameter estimates and significance levels of the welfare indicators are very similar to the values reported in Tables 1 and 2.

The bottom part of each table shows the additional error in the welfare coefficient due to the fact that welfare levels - either poverty or inequality - have been imputed. The first row gives the results obtained when full information about $\mathbf{\Sigma}_{\mathrm{M}}$ can be determined from the unit record data. We use the empirical covariance matrix derived from 100 simulated sets of welfare indicators. Consider first Table 1. Column (1) gives the additional variance - the ('$\widetilde{\mu}$', '$\widetilde{\mu}$') component of the matrix $\beta^2(\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'\Sigma}_{\mathrm{M}}\mathbf{X}(\mathbf{X'}\mathrm{X})^{-1}$. The second column gives the full adjusted variance (8.959 plus column 1) corresponding to a standard error of 3.128. Columns (3) and (4) indicate the share of the model variance in the total variance, and the increase as a percentage of the non-adjusted variance. At over 9% the addition to the variance in the downstream regression coefficient on the headcount, due to the fact that it is estimated, is not trivial. However, the coefficient is still clearly significant.

The next lines in the table explore different ideas for approximating the model covariance matrix $\mathbf{\Sigma}_{\mathrm{M}}$. The results are negative; these simple approximations to the covariance matrix

[11] The available levels of aggregation are (in increasing order of aggregation) household, parroquia, canton, province, and region.

simply do not work, and our quest for a parsimonious approximation to $\mathbf{\Sigma}_{\mathrm{M}}$ will have to continue.

In each case we approximate G, discussed in section 3.2, by taking the share of model error in the variance of the total prediction error in μ, $\mathrm{V_M/V}$, and averaging it over cantons. This gives 0.92 for Rural Costa and 0.66 for Rural Sierra. These numbers are high because idiosyncratic error diminishes in importance at the canton level due to aggregation. Each row makes a different assumption about the degree of correlation, K, between estimates of the expected headcount across cantons within each of the two strata. Clearly in this model using a single value to summarize the correlation leads to underestimation of the model error component - for any value of K between 0 and 1. Note that the underestimation gets *worse* if one allows for more (average) correlation. These results are not general: in a regression without province dummies the approximated model error component *in*creases with correlation and the model error effect is well reproduced for average correlation of $K = 0.15$.

The 'max V' line shows that a crude error approximation (see section 3.2) with ρ equal to the maximum among all prediction error variances, V, gives a safe but rather high upper bound to the model part of the variance in $\widehat{\beta}$. The final lines in the table explore how many terms in a singular value decomposition of $\mathbf{\Sigma}_{\mathrm{M}}$ would be needed to accurately replicate the model error-induced error on the headcount coefficient. Twenty terms suffice, which is some, but not a big gain compared to needing the full $\mathbf{\Sigma}_{\mathrm{M}}$ matrix.

In Table 2 we see that the fact that the welfare variable is imputed makes considerably more difference when it is an indicator of inequality. There are two reasons for this. First, the unadjusted regression results in a much lower significance level for the coefficient on the welfare indicator and second, the prediction error on the inequality measure is much bigger than that of the headcount. On average the prediction (standard) error is 11.7% for the inequality measure and 4.2% for the headcount. Thus, inclusion of the model error in $\widehat{\beta}$ increases its variance by more than 100%. We see that the coefficient on inequality, which appeared to be significant when model error was ignored, is in fact borderline significant at a 10% level (the t-statistic is 1.70). Looking further down the table, we find again that using a single value to summarize the correlation across welfare estimates leads to underestimation of the model error component for any value of K between 0 and 1.[12] However, in this regression the approximation *improves* if one allows for more correlation. The crude error estimation (Max V) gives a very high upper bound in this case and would lead one to (incorrectly) soundly reject a relationship between inequality and garbage collection services.

[12]The table reports results for the (extreme) assumption that all prediction error is model error, or $G = 1$.

4 Endogeneity

In this section we discuss two types of endogeneity issues.

4.1 Endogeneity of W

True welfare W may be correlated with the regression disturbance τ. In this case, one would like to instrument for W, and $\widetilde{\mu}$ may be a better explanatory variable to use in the downstream regression even if W were known.

Example one - Health: Suppose that a health indicator of interest is independent of inequality but both are correlated with an omitted variable "ethnic diversity". Estimating the health regression with true inequality could give a significant, but spurious, coefficient.

Example two - Credit: Suppose that credit availability is independent of poverty but both are correlated with an omitted variable "remoteness". In this situation we would find a negative coefficient on W in a credit regression, but again it would be spurious.

Using $\widetilde{\mu}$ instead of W resolves this type of endogeneity problem.

4.2 Endogeneity of $\widetilde{\mu}$

As in any problem involving instrumental variables, using predicted values may create, rather than resolve, an endogenity problem. However, it is important to realize that when $\widetilde{\mu}$ is correlated with the downstream disturbance τ, the (unknown) true value of welfare, W, would likely also be correlated with the disturbance. There would indeed be an endogeneity problem, but not one special to having used a predicted value for welfare. The usual remedy would apply: instrument $\widetilde{\mu}$.[13] The only cause for additional concern then, would be if by construction $\widetilde{\mu}$ was correlated with τ when W itself was not.

One plausible way to have a regression in which expected welfare is correlated with the disturbance is if one of the variables used in the construction of $\widetilde{\mu}$ should have been included in the downstream regression but is omitted. That said, note that while the suspect variable would have entered the regression, say, linearly, it enters $\widetilde{\mu}$ "mixed" in a non-linear fashion and possibly at a different level of aggregation. So it is not obvious whether the effect of the omitted variable would be picked up on $\widetilde{\mu}$ in the downstream regression.

An investigation of the correlation between selected household-level variables used in the consumption model and the resulting estimates of expected welfare is presented in Table 3.

[13] In principle rather than instrumenting after the fact one could use exogenous variables in the construction of $\widetilde{\mu}$. However, in practice this is unlikely to be feasible because the welfare estimates are typically constructed for targeting purposes. Moreover, appropriate exogenous variables will depend on the particular downstream application.

The first column gives the measure, either the poverty headcount or the GE (0.5) measure of inequality. The second column shows the level of the explanatory variables, i.e. "Parroquia" indicates that the variables are means at that level of aggregation. The third column gives the level of aggregation for the welfare estimate, $\widetilde{\mu}$. The rest of the columns give correlation coefficients between $\widetilde{\mu}$ and the variable indicated in the column heading.

Several points emerge. First, there is far less correlation between the $\widetilde{\mu}$ and other variables when $\widetilde{\mu}$ is an inequality measure. This is not surprising as inequality is particularly non-linear. With a household-level regression, in fact, the "mixing" seems to remove almost all correlation. In household regressions, then, it seems extremely unlikely that including a constructed estimate of inequality will create any endogeneity issues. Second, in many cases we do see considerable correlation. In these situations the best advice would be to instrument $\widetilde{\mu}$. Again, we emphasize that this is not special to using a predicted variable and is likely to be an important precaution even if one were to have true W.

Finally, it is interesting to observe that for both poverty and inequality, the correlations get stronger at higher levels of aggregation. Take eduation of the head, for example. Although estimated poverty at the parroquia level is constructed from household measures of education, it is more strongly correlated (-0.32 vs -0.62) with the average level of education for the parroquia than it is with the household measures used in its construction. There seem to be macro relationships between the variables and the welfare levels that extend beyond their micro relationship with household consumption. These call for further investigation.

5 Predicted Welfare on the Left-Hand Side

We have seen how imputed welfare estimates can be used in a straightforward way as explanatory variables. Many questions of interest in development, however, concern the determinants of distributional outcomes. Exploring these questions requires using imputed variables on the LHS of a regression and on the face of it this looks suspect. The expenditure equation (1) gives a full statistical description of household level consumption. Given the distribution of household observables $\mathbf{z}$ in the target population, and the distribution of the error components η and ε the (expected) distribution of consumption expenditure is fully determined: there seems to be no room for further determination of this distribution. For instance, suppose the expenditure equation involves a household-level education variable. Then it would seem to be very suspect to regress canton-level poverty, imputed from the expenditure equation, on average education in the canton. Since the regression coefficient on average education is completely determined by the expenditure model and the distribution of education in the population; interpreting it as evidence of a direct relationship, at the aggregate level, seems misleading.

5.1 Analysis

For simplicity, let household per-capita expenditure y_{kh} of household h in canton k be determined by the single variable household-level education z_{kh} and an i.i.d. error term u_{kh}, uncorrelated with z_{kh}:

$$\ln y_{kh} = z_{kh} + u_{kh}. \tag{11}$$

The imputed head count at the canton level is[14]

$$\mu_k = \frac{1}{N_k} \sum_{h \in H_k}^{N_k} m_{kh} \Pr(u_{kh} \leq a - z_{kh}),$$

where H_k denotes the set of households in canton k, N_k the total population, and m_{kh} the household size. Obviously, regressing the imputed headcount μ_k on $\overline{z}_k$, the average level of education in location k, will result in a significant regression parameter which seems at best to have only descriptive value. However, we would find essentially the same aggregate relationship if we would have regressed *true* average poverty, W_k, on average education:

$$\begin{aligned} E(W_k|\overline{z}_k) &= E(E(W_k|\{z_{kh}, u_{kh}\})|\overline{z}_k) \\ &= E(\frac{1}{N_k} \sum_{h \in H_k}^{N_k} m_{kh} \Pr(u_{kh} \leq a - z_{kh})|\overline{z}_k) \\ &= E(\mu_k|\overline{z}_k). \end{aligned}$$

The issue is not so much to use an imputed or true variable on the LHS, but to interpret an aggregate relationship as causal or direct: if such a relationship exists, we will find it using either true or imputed variables; if it does not exist, the aggregate fit is a statistical artifact in both cases. Here is our main proposition:

> *If handled carefully, regressions involving imputed indicators of welfare on the LHS and/or the RHS, will give regression coefficients not systematically different from similar regressions, involving the true indicators.*

Note that the above analysis does not hinge on specifying the expenditure model correctly. If the true expenditure generating process differs from the specified expenditure model, the latter's success will simply depend on the degree of correlation of observed variables used in the expenditure regression with the true expenditure-determining variables. But this remains true at the aggregate level, which is equally misspecified or well-specified with true or imputed variables.

[14] For ease of discussion we abstract from model error in this section. Complications from model error can be handled as in the previous sections.

Formally, suppose we want to regress a welfare indicator on explanatory variables z_k, then we have for the imputed and true welfare indicator:

$$E(\mu_k|z_k) = E(E(W_k|\{z_{kh}\})|z_k)$$

and

$$E(W_k|z_k) = E(E(W_k|\{z_{kh}, z_k\})|z_k).$$

Hence, if

$$E(E(W_k|\{z_{kh}\})|z_k) = E(E(W_k|\{z_{kh}, z_k\})|z_k),$$

then

$$E(\mu_k|z_k) = E(W_k|z_k).$$

In other words, if the information in $\{z_{kh}|h = 1, \ldots, N_k\}$ includes the information in z_k, then putting μ_k or W_k on the LHS essentially makes no difference. This condition will be satisfied if z_k is part of the household characteristics $\{z_{kh}\}$ or is otherwise a function of these. More generally, if z_k does not significantly add explanatory power to household per-capita consumption expenditure, beyond the variables z_{kh}, then a regression of W_k on z_k would give essentially the same coefficients as a regression of imputed welfare μ_k on z_k.

Another way to make this point is to consider the regression of W_k on z_k:

$$W_k = z_k\beta + \varepsilon_k. \tag{12}$$

Let $W_k = \mu_k + \omega_k$, with ω_k the (idiosyncratic) prediction error. It follows that

$$\mu_k = z_k\beta + \varepsilon_k - \omega_k. \tag{13}$$

If ω_k is uncorrelated with z_k the latter regression is no more problematic than the former.

Correlation between ω_k and z_k will be negligible if including z_k in the consumption regression (1) does not lead to significant improvement of the fit. This will be the case if the z variables are constructed from census data, or more generally from the same data sources used in the construction of the welfare indicators. These variables, if not included already, will have been considered for inclusion in the consumption regression so that correlation between ω_k and z_k is unlikely to be a problem.

On the other hand if one has (location) data z_k from other sources and there is no practical way to test how well it would have performed as an additional explanatory variable in the consumption regression, then correlation between z_k and ω_k in equation (13) might compromise

the estimation of β. A solution for this would be to instrument z_k with census data.[15]

Finally, a household-level statistical relationship such as the expenditure equation (1) does not preclude the existence of aggregate causal relationships. The expenditure model and the information on the distribution of explanatory variables in the population (from the census) do allow one to predict statistical relationships at aggregated levels. But as emphasized in Elbers, *et al.,* (2003, p. 356) the parameters of the expenditure model measure correlation not causality. The predicted aggregate relationships are based on these correlations and therefore say nothing about the existence or non-existence of causal aggregate relationships. The correlation patterns found at the household level in the survey and census data could very well have sprung from an aggregate causal relationship. As always in regressions: *caveat emptor*. It takes meticulous diagnostics before a regression coefficient can be interpreted as marginal impact. The use of imputed rather than true variables does not in any way simplify or compound that basic difficulty.

5.2 Example

Consider a Kuznets-type regression of v_k, the variance of log per-capita household consumption in location k on average consumption $\bar{y}_k$, both estimated using the model in equation (11). We take the distribution of both the education variable z_{kh} and the error term u_{kh} to be normal. Hence we find

$$\begin{aligned} v_k &= \text{var}(z_{kh}) + \text{var}(u_{kh}) \\ \bar{y}_k &= e^{\overline{z}_k + \frac{1}{2}v_k}. \end{aligned}$$

Assume that both $\text{var}(z_{kh})$ and $\text{var}(u_{kh})$ are heteroskedastic; for the sake of argument, let both depend on the average level of education:

$$v_k = \text{var}(z_{kh}) + \text{var}(u_{kh}) = \varphi(\overline{z}_k).$$

Differentiating, we find

$$\begin{aligned} dv_k &= \varphi'(\overline{z}_k) d\overline{z}_k \\ d\bar{y}_k &= \bar{y}_k(1 + \frac{1}{2}\varphi'(\overline{z}_k)) d\overline{z}_k. \end{aligned}$$

[15] Such instrumenting requires access to census data. However, the target regression will typically not be at the household level but at higher levels of aggregation for which it may be easier to obtain the necessary census-based data.

Hence,

$$\frac{dv_k}{d\bar{y}_k} = \frac{\varphi'(\overline{z}_k)}{\bar{y}_k(1 + \frac{1}{2}\varphi'(\overline{z}_k))}.$$

The slope of the Kuznets curve is ultimately determined by the heteroskedasticity function $\varphi(\overline{z}_k)$. Here we have calculated the slope using imputed variables. The main point to note is that explanation of the Kuznets curve depends on explanation of the function $\varphi(\overline{z}_k)$, which itself has nothing to do with using imputed or true variables. If the use of imputed variables has helped to obtain more information for the analysis of $\varphi(\overline{z}_k)$, that is only an improvement.

6 Conclusions

Some of the oldest research activities in Development Economics involve the analysis of distributional indicators in relation to other indicators. The Kuznets curve, relating income inequality to average income level, is a famous example. Another example is the never-ending debate on the relationship between inequality and growth, with disagreement both on the sign of the relationship and the direction of causality. One of the main motives behind our poverty mapping project was to compile more disaggregate and closely comparable estimates of distributional measures to begin building a better empirical foundation for these discussions.

Because the estimated inequality and poverty measures are predicted values rather than data, their use in regression analysis requires attention to econometric issues. We have discussed how imputed distributional indicators can be used as explanatory variables in regressions. Our conclusion is that imputed variables on the right-hand side can be regarded as a special kind of instrumented variables and, if handled correctly, can be safely used in estimation. This is demonstrated in regressions using data from Ecuador. In a canton-level regression of garbage collection on imputed headcount poverty, the fact that explanatory variables were imputed had a small but non-negligible effect on the estimated standard errors of the regression coefficients. On the other hand, in a similar regression on local inequality the increase in error due to imputation was far greater.

To calculate correct standard errors requires knowledge of the model error in the welfare estimates used as explanatory variables. Our (limited) experience suggests that there may be no simple parsimonious subsitute for the full covariance matrix of model errors. This need not imply, however, that only those with access to census record data will be able to proceed. Those calculating the welfare estimates can store the requisite information for use by downstream researchers, along side the point estimates and their prediction errors. The most efficient way to store the information, whether as matrix $\widehat{\mathbf{\Sigma}}_{\mathrm{M}}$, as vectors of simulated draws $\widetilde{\mu}^r$ (step 3 in section 3.1), or some other form, would depend on the context.

Using imputed variables on the left-hand side is trickier, but essentially such regressions

yield results no different from what would follow from similar regressions involving the true welfare indicators. However, such regressions might suffer from problems of omitted variable bias inherent in using imputed variables. We have discussed ways to avoid such problems.

We conclude that the scope for analysis of distributional issues at various levels of aggregation is vastly expanded by the availability of poverty maps.

References

[1] Demombynes, Gabriel, Chris Elbers, Jenny Lanjouw, Peter Lanjouw, Johan Mistiaen and Berk Ozler (2003) "Producing an Improved Geographic Profile of Poverty: Methodology and Evidence from Three Developing Countries," WIDER Discussion Paper no. 2002/39. Forthcoming in Rolph van der Hoeven and Anthony Shorrocks (eds.) *Growth, Inequality and Poverty.* (Oxford: Oxford University Press).

[2] Elbers, Chris, J.O. Lanjouw and Peter Lanjouw. (2003) "Micro-Level Estimation of Poverty and Inequality," *Econometrica.* Vol. 71, no. 1, pp. 355-64.

[3] _____(2002)."Micro-Level Estimation of Welfare," Policy Research Working Paper no. WPS 2911. The World Bank.

[4] Greene, William H. (2000) *Econometric Analysis*. Fourth Edition. (New Jersey: Prentice-Hall, Inc.)

[5] Murphy, Kevin M. and Robert H. Topel (1985) "Estimation and Inference in Two-Step Econometric Models," *Journal of Business & Economic Statistics*, Vol. 3, no. 4, pp. 370-79.

Model Variance in Downstream Regression Coefficients and Approximations
Headcount and Canton-level Data

Standard Regression Output

Coefficient on population	0.332
Coefficient on the headcount, $\hat{\beta}$	-19.132
Estimated (robust) standard error of $\hat{\beta}$	2.993
Estimated variance of $\hat{\beta}$	8.959
Adjusted R^2	0.66

Analysis of Estimated Model Variance in $\hat{\beta}$

	Model variance in $\hat{\beta}$ (1)	Total variance in $\hat{\beta}$ (2)	Model share (1)/(2) (3)	Percentage increase in variance (4)
Using 'True' Σ_M	0.826	9.786	0.084	9.22
K-values				
0.00	0.416	9.375	0.044	4.64
0.33	0.366	9.325	0.039	4.08
0.66	0.315	9.274	0.034	3.52
1.00	0.263	9.222	0.029	2.93
Max V	1.971	10.930	0.180	22.00
Single Value Decomposition				
5 terms	0.200	9.159	0.022	2.23
10 terms	0.520	9.479	0.055	5.80
15 terms	0.759	9.718	0.078	8.47
20 terms	0.803	9.762	0.082	8.96

Table 1. The effect of prediction error in explanatory variables in a regression of an index of garbage collection on imputed headcount poverty. Source: authors' calculations.

Model Variance in Downstream Regression Coefficients and Approximations
GE(0.5) Inequality Measure and Canton-level Data

Standard Regression Output

Coefficient on population	0.413
Coefficient on the headcount, $\hat{\beta}$	11.951
Estimated (robust) standard error of $\hat{\beta}$	4.876
Estimated variance of $\hat{\beta}$	23.771
Adjusted R^2	0.52

Analysis of Estimated Model Variance in $\hat{\beta}$

	Model variance in $\hat{\beta}$ (1)	Total variance in $\hat{\beta}$ (2)	Model share (1)/(2) (3)	Percentage increase in variance (4)
Using 'True' Σ_M	25.933	49.704	0.522	109.10
K-values				
0.00	8.664	32.435	0.267	36.45
0.33	13.290	37.060	0.359	55.91
0.66	17.915	41.685	0.430	75.37
1.00	22.680	46.451	0.488	95.41
Max V	36.230	60.000	0.604	152.41
Single Value Decomposition				
5 terms	25.719	49.490	0.520	108.20
10 terms	25.827	49.598	0.521	108.65
15 terms	25.830	49.601	0.521	108.66
20 terms	25.849	49.620	0.521	108.75

Table 2. The effect of prediction error in explanatory variables in a regression of an index of garbage collection on imputed GE(0.5) inequality. Source: authors' calculations.

Measure	Regression	$\tilde{\mu}$	Education household head	Age of household head	Household head has no spouse	Indigenous language spoken in household	Sole use sewage connection	Shared use sewage connection	
Headcount	Household	Household	-0.36	<0.01	-0.33	0.14	-0.43	-0.06	0.07
		Parroquia	-0.32	0.05	0.01	0.26	-0.34	-0.17	0.07
		Canton	0.20	0.05	<0.01	0.21	-0.22	-0.11	0.02
	Parroquia	Parroquia	-0.62	0.23	0.02	0.32	-0.71	-0.51	-0.03
	Canton	Canton	-0.69	0.40	-0.08	0.38	-0.78	-0.57	0.05
GE (0.5)	Household	Parroquia	0.11	<0.01	0.02	0.05	0.06	0.07	-0.11
		Canton	0.07	0.01	0.04	0.10	-0.02	0.05	-0.12
	Parroquia	Parroquia	0.15	0.01	0.17	0.11	0.06	0.13	-0.27
	Canton	Canton	0.08	0.07	0.24	0.15	-0.08	0.19	-0.37

Table 3. Correlations between welfare indicators and household characteristics, used in their construction. Source: authors' calculations using unit records of Ecuador population census, 1990.